

Link-Based and Content-Based Evidential Information in a Belief Network Model

Ilmério Silva^{1,2} Berthier Ribeiro-Neto¹ Pável Calado¹ Edleno Moura^{1,3} Nívio Ziviani¹

¹Universidade Federal de Minas Gerais
30.123-970 Belo Horizonte-MG, Brazil
{ilmerio,berthier,pavel,edleno,nivio}@dcc.ufmg.br

²Universidade Federal de Uberlândia
38.408-100 Uberlândia-MG, Brazil
ilmerio@ufu.br

³Universidade do Amazonas
69.077-000 Manaus-AM, Brazil
edleno@dcc.fua.br

This work was supported in part by Finep/MCT/CNPq Grant 76.97.1016.00, project SIAM under program Pronex, and in part by CNPq individual grants 300.188/95-1 and 520916/94-8.

Abstract This work presents an information retrieval model developed to deal with hyperlinked environments. The model is based on belief networks and provides a framework for combining information extracted from the content of the documents with information derived from cross-references among the documents. The information extracted from the content of the documents is based on statistics regarding the keywords in the collection and is one of the basis for traditional information retrieval (IR) ranking algorithms. The information derived from cross-references among the documents is based on link references in a hyperlinked environment and has received increased attention lately due to the success of the Web. We discuss a set of strategies for combining these two types of sources of evidential information and experiment with them using a reference collection extracted from the Web. The results show that this type of combination can improve the retrieval performance without requiring any extra information from the users at query time. In our experiments, the improvements reach up to 59% in terms of average precision figures.

Keywords : IR models, exploiting hyperlinked structure, content-based retrieval

1 Introduction

Two main strategies have been used to determine the ranking of documents in Web search engines. One is to use document keywords as indexing elements and compare them against the user query. In this case, only the document *content* is used as a source of evidence to determine its ranking score [17]. Another alternative is to use the knowledge derived from the link structure in the

Web to determine the ranking of the documents for each user query [4, 12]. In this case, the ranking score of a document is determined in terms of the number of other pages or documents that contain references to (or from) it.

Recent studies have suggested that combining link-based and content-based evidential information can improve the retrieval performance of Web search services [2, 5]. The main reasons seem to be that most queries on Web search services are short and imprecise [11] and that the users themselves are frequently uncertain about what exactly they are looking for. Therefore, the task of finding relevant information in Web search engines is hard and the use of any source of additional evidence available in this environment must be considered in an attempt to improve the retrieval results.

In this work we study new alternatives for combining link-based and content-based evidential information with the objective of improving the quality of Web search engines results. We adopt the Bayesian belief networks [13] as a unifying framework that naturally allows representing and combining both types of evidential knowledge in a single information retrieval model. We present basic alternatives for such combination in our belief network model. Through experimentation with a reference collection extracted from the Web, we study and compare these alternatives. We explicitly compare the results of a combined ranking with those yielded by a vectorial ranking (which uses only content-based evidence). The experiments indicate that this type of combination can improve the retrieval performance considerably without requiring any extra information from the users at query time.

2 Related Work

Turtle and Croft [19] were the first ones to propose the use of Bayesian networks to model information retrieval problems. In their work, they represent queries and documents on an inference network. Following, Ribeiro-Neto and Muntz [15] have used Bayesian networks to model evidence derived from past queries and to combine it with the vector space model [17]. In both cases, the formalism of Bayesian networks provided a sound framework for representing, quantifying, and combining two or more sources of evidence in support of a ranking for the documents in the answer set. In here, we explore the use of

Bayesian networks to represent and combine link-based and content-based (or keyword-based) evidential information.

Keyword-based evidential information is normally quantified using statistics on the occurrence of keywords on the documents and constitutes one of the most fundamental metrics for ranking in information retrieval systems. The idea is to use the keywords and statistics on their occurrences to determine which documents have content more similar to the user query. An alternative and complementary approach is the use of relationships among the documents, such as cross-referencing and bibliographic citation [3, 8, 18]. Cross-referencing can be implemented as links in a hyperlinked environment and has recently been used to compute the ranking in Web information systems [2, 4, 5, 12]. Brin and Page [4] propose an algorithm that uses the link structure to indicate how *authoritative* is a document with regard to a search topic or user query. The authoritative degree can be interpreted as a popularity measure of the document based on the link structure surrounding it. Kleinberg [12] proposes a new algorithm that also analyzes the link structure. This algorithm finds pages, called *hubs*, that point to many other pages and uses them as a component to determine the *authoritative* degree of the documents. Documents pointed by many hubs take higher *authoritative* degree. Kleinberg's algorithm can be restricted to the subset of documents that compose the answer set to a user query. This subset is usually called the *local* set of documents and constitutes a major focus of our study.

Previous works have proposed combining link-based (link structure) with keyword-based (content) pieces of evidence in a single information retrieval model. The algorithm of Chakrabarti et al [5] combines the local link analysis described in [12] with keyword-based evidence. Their system uses the text surrounding the links as keyword-based evidence to determine a weight for each link analyzed. After determining the weights of each link, the ranking is computed using a weighted version of the algorithm proposed by Kleinberg [12].

Bharat and Henzinger [2] have also studied alternatives to combine link analysis with keyword-based evidence. Their work has two main differences when compared with the algorithm of Chakrabarti et al [5]. First, they also use keywords to determine the relevance of the links, but this is done taking all the words in the document. Second, they expand the original query using the keywords of the documents in the local answer set and compute the weight of each link based on the expanded query. This expansion process improves the retrieval performance but can be somewhat expensive because it greatly increases the number of terms to be processed. Bharat and Henzinger have also modified the algorithm proposed by Kleinberg [12] using heuristics to reduce the weight of some links that degrade the results. An example of such type of links occurs when a set of links from a same site all point to a unique page.

In this paper we present new ways to combine keyword-based evidence and link analysis. Our work differs from previous studies in several directions. First, we adopt Bayesian networks as a unifying modeling framework. This provides a formal approach to the problem and also produces flexible models which can be easily modified to include additional sources of evidences. Second, taking advantage of this increased flexibility, we use link-based evidence to change the ordering of the documents generated by a vectorial ranking (computed using keyword-based evidence), while previous works have used keyword-based evidence only to change the link weights at the phase of link analysis. An extra advantage of our model is that it does not require any query expansion, which improves its overall efficiency. Further, we combine informa-

tion on both authorities and hubs with content-based evidential information. This new combination gives better retrieval results than strategies that use either authorities or hubs in combination with keyword-based evidence.

3 Link-Based Evidence

One of the richest sources of information in a hyperlinked environment, such as the Web, is the knowledge about its link structure. In fact, such knowledge frequently encodes human judgment about the documents which can be of critical importance in the generation of a good ranking. Kleinberg [12] uses this information to measure the importance of a document based on two metrics: a degree of *authority* and a degree of *hub*. A good *authority* is defined as a document with a high number of incoming links from good *hubs*. Recursively, a good *hub* is defined as a document with a high number of out-coming links that point to good *authorities*. Kleinberg has also proposed an algorithm to compute a degree of goodness for hubs and authorities based on an analysis of the link structure surrounding the documents in the local answer set to a user query. In this case, we say that the algorithm computes a *local authority* for each document. When the algorithm is applied to the whole set of documents in the collection, we say that it computes a *global authority* for each document.

Computing a Degree of Local Hub and Local Authority

We interpret a collection of hyperlinked documents as a directed graph \mathcal{G} where each document (page) is represented by a node of \mathcal{G} and each link between two documents is represented by a directed edge of \mathcal{G} . By assumption, a link from a document D to another document D' implies that the author of the document D endorses the document D' .

Consider a collection of hyperlinked documents and its associated directed graph \mathcal{G} . Given a user query Q , the *local hub* and *local authority* values of each document can be computed using the link structure associated with the documents in the local answer set. Let $\mathcal{Q} = (\mathbf{V}, \mathbf{E})$ be a subgraph of \mathcal{G} such that each node of \mathbf{V} represents a document related to the query Q and the set of edges \mathbf{E} represents a set of links related to the documents in \mathbf{V} . The subgraph \mathcal{Q} is computed by the algorithm A in Figure 1, while the local hub and local authority values of each document are computed by the algorithm B in Figure 2. These two algorithms were proposed by Kleinberg in [12]. Details about the convergence of the algorithm B can also be found in [12].

Computing a Degree of Global Hub and Authority

The algorithm B can be applied to the graph \mathcal{G} representing all the documents and links in a collection of hyperlinked documents (instead of a subgraph derived from the documents related to a topic of interest). In this case, the hub and authority values computed are referred to as degrees of *global hub* and of *global authority*.

The ideas described here can be easily extended to include global hub and authority values in the model. However, our preliminary experiments have shown that global values bring little information to produce a useful ranking in the reference collection we used (almost all pages having zero values). The algorithm can possibly

Algorithm A (Q, T, L)

Q : a query
 T : the number of documents to be considered in an initial local answer set
 L : a limit on the number of parent documents (of a document) to be considered
Let \mathbf{V}_0 be the set of the T top documents in the ranking generated by the query Q , using any ranking algorithm
Let $\mathbf{V} := \mathbf{V}_0$
For each document $D \in \mathbf{V}_0$ Do
 Let \mathbf{C} (children of D) be the set of all the documents pointed by D
 Let \mathbf{P} (parents of D) be the set of all the documents that point to D
 If \mathbf{P} contains more than L documents Do
 Let \mathbf{P}_l be a subset of the L documents arbitrarily selected from \mathbf{P}
 Let $\mathbf{P} := \mathbf{P}_l$
 end
 Let $\mathbf{V} := \mathbf{V} \cup \mathbf{C} \cup \mathbf{P}$
end
Let \mathbf{E} be the set of all the edges from the document D_i to the document D_j , where $D_i \in \mathbf{V}$ and $D_j \in \mathbf{V}$
Return (\mathbf{V}, \mathbf{E})

Figure 1: Algorithm A, for computing a subgraph (\mathbf{V}, \mathbf{E}) related to a given query Q .

Algorithm B(\mathbf{V}, \mathbf{E})

\mathbf{V} : a set of documents
 \mathbf{E} : a set of directed edges linking documents of \mathbf{V}
Let N be the number of documents in \mathbf{V}
Let $\mathbf{X} := (X_1, X_2, \dots, X_N)$ be a vector of N values for authorities, all initially set to 1
Let $\mathbf{Y} := (Y_1, Y_2, \dots, Y_N)$ be a vector of N values for hubs, all initially set to 1
While the vectors \mathbf{X} and \mathbf{Y} have not converged Do
 For $i := 1$ to N Do
 $X_i := \sum_{(D_j, D_i) \in \mathbf{E}} Y_j$ end
 For $i := 1$ to N Do
 $Y_i := \sum_{(D_i, D_j) \in \mathbf{E}} X_j$ end
 Normalize the vectors \mathbf{X} and \mathbf{Y} such that $\sum_i X_i^2 = \sum_i Y_i^2 = 1$
end
Return \mathbf{X} and \mathbf{Y} in descending order of their values

Figure 2: Algorithm B, for computing the authority and the hub values of each node (or document).

be refined to avoid this problem, but we leave these refinements for future work and focus here only on local values.

4 The Belief Network Model for IR

This section shows how to model a content-based solution to the information retrieval problem using Bayesian networks. For this task, we adopt the *belief network model* defined in [15]. This model takes an epistemological view (as opposed to a frequentist view) of the information retrieval problem and interprets probabilities as degrees of belief devoid of experimentation, as also done in [19, 21]. This is the reason for calling it a *belief network model*.

The belief network model adopts Bayesian networks as its basic foundation. Bayesian networks are useful because they provide a graphical formalism for explicitly representing independencies among the variables of a joint probability distribution. The probability distribution is represented through a directed acyclic graph whose nodes represent the random variables of the distribution. The relationships among these variables are modeled as directed edges (in the graph) which represent causal dependencies among the linked variables (or nodes)¹. The strengths of these dependencies are expressed by conditional probabilities. The fundamental principle is that the known independencies among random variables of a domain are declared explicitly and that a joint probability distribution is synthesized from this set of declared independencies.

In a traditional content-based information retrieval system, the documents and the user queries are usually represented as sets of keywords. As a result of this interpretation, queries and documents are treated analogously as proposed in [15]. Figure 3 illustrates a belief network which reflects this symmetry. In this network, each node

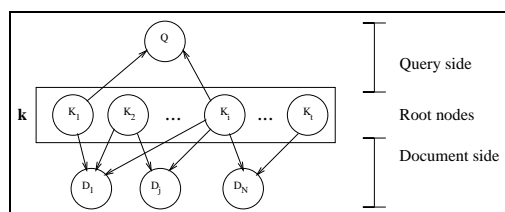


Figure 3: Belief network for a query Q composed of the keywords K_1 and K_i .

D_j models a document D_j , the node Q models the user query Q , and the K_i nodes model the keywords in the collection. The vector \mathbf{k} is used to refer to any of the possible states of the *root nodes* K_i (these are the nodes without parents). A *binary* random variable is associated to the node Q , which is also denoted by Q . In this notation it should always be clear whether we are referring to the query, to the node in the network, or to its associated binary variable. The variable Q is 1, denoted by q , to indicate that Q is active and $Q = 0$, denoted by \bar{q} , to indicate that Q is inactive. Analogously, a *binary* random variable D_j is associated with the document node D_j . The variable D_j is 1, denoted by d_j , to indicate that D_j is active and $D_j = 0$, denoted by \bar{d}_j , to indicate that the variable D_j is inactive. A *binary* random variable K_i is also associated with each keyword K_i . All of these

¹Although an edge linking a node Y to a node X is frequently used to express that Y causes X , this interpretation of edges in Bayesian networks is not the only one possible.

variables are binary because this is simple and provides enough semantics for modeling the information retrieval problem. Varying degrees of relevance are represented in the model as conditional probabilities, as we later discuss.

Instantiation of the root nodes *separates* the document nodes from the query node, making them mutually independent (see Bayesian theory for more details [13]). Thus, in the belief network of Figure 3, we say that the query is on the *query side* of the network, while the documents are on the *document side* of the network.

In the network of Figure 3, the ranking computation is based on quantifying the similarity between a document D_j and the query Q by the probability $P(D_j = 1|Q = 1)$, or simply $P(d_j|q)$ (i.e., the probability that the variable D_j is active given that the variable Q is active). By the rule of total probabilities and the independencies modeled in the network we can write,

$$P(d_j|q) = \eta \sum_{\mathbf{k}} P(d_j|\mathbf{k}) P(q|\mathbf{k}) P(\mathbf{k}) \quad (1)$$

where η is a normalizing constant [13]. This is the generic expression for computing the rank of a document D_j with regard to the query Q , in our belief network model.

Modeling the Vector Space Model

In [14], it is demonstrated that Eq. (1) can be used to represent any of the classic models in IR namely, the Boolean, the vector, and the probabilistic models, and also to represent any ranking generated by the inference network model [19]. Here, we review how to use a belief network to compute a ranking generated by the vector space model [1, 16, 20].

To compute a vectorial ranking in our belief network, we specify the probabilities $P(\mathbf{k})$, $P(q|\mathbf{k})$ and $P(d_j|\mathbf{k})$. First, we define the prior probabilities $P(\mathbf{k})$ associated with the root nodes, as follows

$$P(\mathbf{k}) = \begin{cases} 1 & \text{if } \forall_i g_i(\mathbf{q}) = g_i(\mathbf{k}) \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

where $g_i(\mathbf{u})$ is a function that returns the value of the i^{th} variable in the vector \mathbf{u} . Eq. (2) establishes that the only state \mathbf{k} of the set \mathbf{K} of root nodes which is taken into account is exactly that one for which the active keywords are exactly those in the query Q .

For $P(q|\mathbf{k})$, we write

$$P(q|\mathbf{k}) = \begin{cases} 1 & \text{if } \forall_i g_i(\mathbf{q}) = g_i(\mathbf{k}) \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

For $P(d_j|\mathbf{k})$ we write

$$P(d_j|\mathbf{k}) = \frac{\sum_{i=1}^t w_{ij} \cdot w_{ik}}{\sqrt{\sum_{i=1}^t w_{ij}^2} \sqrt{\sum_{i=1}^t w_{ik}^2}} \quad (4)$$

where w_{ik} and w_{ij} are *tf-idf* weights [1, 16] used in the vector model². By substituting Eqs. (2) to (4) into Eq. (1), we obtain a ranking for the D_j documents, expressed as $P(d_j|q)$, which preserves the ordering dictated by a vectorial ranking.

²This specification is valid and consistent because $P(d_j|\mathbf{k})$ measures the cosine of the angle between two vectors, which is a number between 0 and 1.

5 Modeling Link-Based Evidence on a Belief Network

We expand now the belief network model discussed above to also include evidences extracted from the link structure of the environment. This is accomplished by adding new edges, nodes, and probabilities to the original network presented in Figure 3. We say that this expansion is modular in the sense that it preserves all the properties of the previous network. Furthermore, this strategy allows us to combine the keyword-based evidence (which summarizes semantic knowledge on content) associated with the vector space model with link-based evidential knowledge (which summarizes semantic knowledge on document relationships), in a natural and convenient way.

In Figure 4, the left hand side of the network represents the original network of Figure 3 with the following adaptations: each document node D_j is renamed as Dc_j (for content-based). The right hand side of the network models the link-based *local* sources of evidence, which are obtained from the link structure associated with the set of documents in the answer set to a query (frequently referred to as the *local set* of documents).

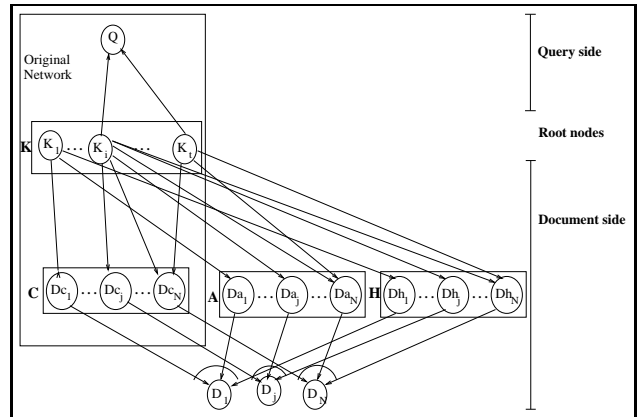


Figure 4: Bayesian network expanded with link-based evidence.

To represent link-based evidential knowledge in the network, we associate two new nodes Dh_j and Da_j with each document D_j in the local set of documents. The node Dh_j , with which we associate a binary random variable also named Dh_j , models evidence associated with the document D_j as a hub. This evidence is computed from the link structure associated with the local set of documents related to the query Q (see Section 3) and is represented in our network as the conditional probability of Dh_j being active given the keywords in the query Q (and given an implicit knowledge of the local link structure). Analogously, we associate a binary random variable Da_j with the node Da_j to model evidence associated with the document D_j as an authority. Thus, we now have three sets of nodes representing evidential knowledge associated with the documents in the network. The set \mathbf{H} , which contains the nodes representing hub evidence; the set \mathbf{A} , which contains the nodes representing authority evidence; and the set \mathbf{C} , whose nodes represent content based evidence. The state of the associated random variables is given by \mathbf{h} , \mathbf{a} , and \mathbf{c} , respectively.

The set of nodes \mathbf{K} is used to model the occurrence of keywords in the query Q and, once instantiated, induces beliefs on each of the nodes in the sets \mathbf{C} , \mathbf{H} and \mathbf{A} . The propagation of these beliefs in the network is done according to the conditional probabilities governing

the relationships between the set \mathbf{K} and each of the sets \mathbf{C} , \mathbf{H} , and \mathbf{A} . These conditional probabilities are specified based on the vector space model and on Kleinberg's algorithm, as we later discuss.

With each node Dh_j of \mathbf{H} is associated a binary random variable Dh_j . This variable is 1 to indicate that the local hub evidence associated with the document D_j is to be considered in the ranking computation. Also, with each node Da_j of \mathbf{A} is associated a binary random variable Da_j , which is 1 to indicate that the local authority evidence associated with the document D_j is to be considered in the ranking computation. The node D_j represents the combination of content-based and link-based evidential knowledge from the left and right hand sides of the network. The conditional probabilities, discussed below, define how these evidences are combined.

General Equation for Ranking Computation

In Figure 4, the rank $P(d_j|q)$ associated with a document D_j can be computed using Eq. (1). However, the conditional probability $P(d_j|\mathbf{k})$ now depends on link-based and content-based pieces of evidence which have to be combined through a disjunctive operator *or*. This is accomplished as follows:

$$P(d_j|\mathbf{k}) = 1 - (1 - P(dc_j|\mathbf{k})) \times (1 - P(dh_j|\mathbf{k})) \times (1 - P(da_j|\mathbf{k})) \quad (5)$$

Substituting the Eq. (5) into Eq. (1), we can write

$$P(d_j|q) = \eta \sum_{\mathbf{k}} [1 - (1 - P(dc_j|\mathbf{k})) \times (1 - P(dh_j|\mathbf{k})) \times (1 - P(da_j|\mathbf{k}))] \times P(q|\mathbf{k}) \times P(\mathbf{k}) \quad (6)$$

The computation of the probability $P(d_j|\mathbf{k})$ depends on the states of the nodes Dc_j , Da_j , and Dh_j . The probability $P(q|\mathbf{k})$ can be computed using the states of the root nodes K_i . Through the proper specification of the states of all these nodes, we can establish interesting alternatives for computing the rank of a document D_j with regard to a query Q .

6 Ranking Computation

As discussed in Section 4, the belief network model can represent the vector model through proper specification of the conditional probabilities in the network. To simplify our notation, let R_{jq} be a reference to the vectorial rank of the document D_j with regard to a query Q computed according to our network model using Eq. (4). Thus,

$$R_{jq} = \frac{\sum_{i=1}^t w_{ij} \cdot w_{iq}}{\sqrt{\sum_{i=1}^t w_{ij}^2} \sqrt{\sum_{i=1}^t w_{iq}^2}} \quad (7)$$

Case 1: Content-Based Ranking

For representing a ranking based solely on document content, we ignore (for now) the knowledge derived from the local link structure. This is accomplished in our network model by defining

$$P(dh_j|\mathbf{k}) = 0 \quad (8)$$

$$P(da_j|\mathbf{k}) = 0 \quad (9)$$

Applying equations (2), (3), (4), (7), (8), and (9) to Eq. (6), we obtain

$$P(d_j|q) = \eta \times R_{jq} \quad (10)$$

Therefore, the general network of Figure (4) naturally subsumes a ranking dictated by the vector space model.

Case 2: Ranking Based on Hub Evidential Knowledge

To represent a ranking that depends only on link-based knowledge, we redefine the conditional probability $P(dc_j|\mathbf{k})$ as

$$P(dc_j|\mathbf{k}) = 0 \quad (11)$$

which allows ignoring (for now) evidence associated with a content-based ranking.

Following, we define the believes associated with the knowledge that comes from the local link structure. Let H_{jq} be the local hub value computed by Kleinberg's algorithm (see Section 3) for a document D_j , with regard to a query Q . In our belief network, the local hub evidence associated with D_j is modeled as a conditional probability attached to the node Dh_j . We can then write

$$P(dh_j|\mathbf{k}) = H_{jq} \quad (12)$$

where H_{jq} is a normalized version of the local hub value of the document D_j with regard to the query Q . To exclude information on authoritative knowledge, we use Eq. (9).

Applying equations (2), (3), (9), (11), (12) into Eq. (6) we obtain

$$P(d_j|q) = \eta \times H_{jq} \quad (13)$$

In this case, our network simply reproduces a ranking based on local hub values.

Case 3: Ranking Based on Authority Evidential Knowledge

As in Case 2 above, we define $P(dc_j|\mathbf{k}) = 0$. Further, let L_{jq} be the local authority value computed by Kleinberg's algorithm. In our belief network, the local authority of a document D_j is modeled as a conditional probability attached to the node Da_j . We can then write

$$P(da_j|\mathbf{k}) = L_{jq} \quad (14)$$

where L_{jq} is a normalized version of the local authority value of the document D_j with regard to the query Q . To exclude information on hub knowledge, we use Eq. (8).

Applying Eqs. (2), (3), (8), (11), and (14) into Eq. (6) we obtain

$$P(d_j|q) = \eta \times L_{jq} \quad (15)$$

In this case, our network simply reproduces a ranking based on local authority values.

Case 4: Combining Content-Based and Hub-Based Pieces of Evidence

Using Eq. (4) instead of Eq. (11) and the concise notation in Eq. (7), we now discuss how our network model can be used to naturally combine keyword-based evidential knowledge with link-based evidential knowledge.

Applying Eqs. (2), (3), (4), (7), (9), and (12) into Eq. (6) we obtain

$$P(d_j|q) = \eta \times [1 - (1 - R_{jq}) \times (1 - H_{jq})] \quad (16)$$

Case 5: Combining Content-Based and Authority-Based Pieces of Evidence

Applying Eqs. (2), (3), (4), (7), (8), and (14) into Eq. (6) we obtain

$$P(d_j|q) = \eta \times [1 - (1 - R_{jq}) \times (1 - L_{jq})] \quad (17)$$

Case 6: Combining Content-Based, Hub-Based and Authority-Based Pieces of Evidence

Applying equations (2), (3), (4), (7), (12), and (14) into Eq. (6), we obtain

$$P(d_j|q) = \eta \times [1 - (1 - R_{jq}) \times (1 - H_{jq}) \times (1 - L_{jq})] \quad (18)$$

Summary of Ranking Alternatives

Table 1 summarizes the six alternative rankings modeled in our network.

7 Evaluation

We first present the reference collection (composed of Web pages) we used in our experiments. Following, we discuss our results.

The Reference Collection

The reference collection is composed of a database of Web pages, a set of example Web queries, and a set of relevant documents associated with each example query, as we now describe.

The database is composed by 3,027,540 pages of the Brazilian Web (domain .br). The pages were automatically collected by the document collector CoBWeb, described in [6], and indexed using inverted lists [7]. Some characteristics of the database used are summarized in Table 2.

The 20 example queries were selected among the set of most frequently asked queries in the *ToDoBR* search engine (<http://www.todobr.com.br>). This search engine was launched by us in 1999 and is the most complete search engine in the Brazilian scenario. To select the queries for our experiments, we used a log with 100,000 queries submitted to *ToDoBR*. Actually, some frequent queries related to sex were not considered. The mean number of keywords per query is 1.6, as shown in Table 2.

For each of our 20 example queries, we compose a query pool formed by the top 10 documents (or pages) generated by each of our 6 types of network ranking. Thus, each query pool contains at most 60 distinct pages. The average number of pages per query pool is 38.15. All documents in each query pool were submitted to a manual evaluation. The average number of relevant pages (or documents) per query pool is 17.05. The pooling method used here is the same as the method used by the Web-based collection of Trec [9, 10].

Results

Figure 5 illustrates the retrieval performance, in terms of precision-recall figures, for the first three case rankings presented in Section 6: vector, hub, and authority. We observe that the vector ranking is always superior for our reference collection. This shows the strength of traditional IR ranking techniques. While this is not conclusive in terms of the whole Web, it clearly indicates that content-based ranking must be always considered. We also notice that a hub-based ranking yields quite good results in the case of our test collection.

In Figure 6, we investigate the impact of combining the vector and authority rankings in our belief network model. The results indicate that this combination yields

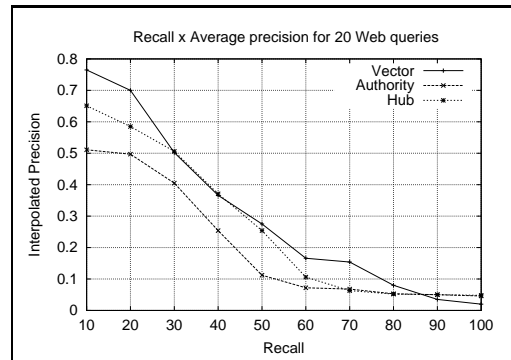


Figure 5: Average precision figures for vector, authoritative, and hub rankings.

precision figures which are always superior to those provided by each ranking in isolation. Particularly, the authority ranking contributes to improve the overall precision for middle and high recall levels (where the vector ranking is not as good as it is at low recall levels).

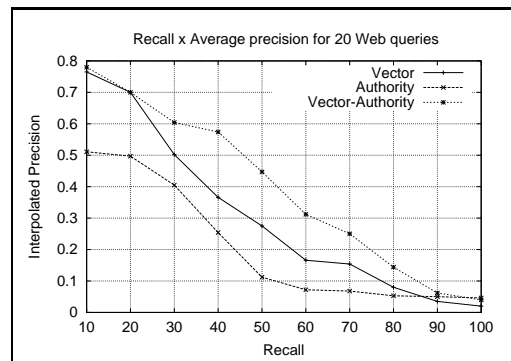


Figure 6: Average precision figures for vector, authority, and vector-authority network rankings.

In Figure 7, we investigate the impact of combining the vector and hub rankings in our belief network model. Again we observe that this combination always yields higher precision figures than those obtained by each ranking in isolation.

Finally, in Figure 8, we investigate the impact of combining the vector, authority, and hub rankings. Except for a slim decrease in precision at very low recall levels, this three-way combination of evidences yields superior results. The belief network model is able to take advantage of the distinct nature of each of our three types of evidential knowledge to provide improved overall retrieval performance. This is an interesting and new result which indicates the strength of belief networks as a framework for consistently combining distinct pieces of evidence on support of a relevance ranking (a characteristic also observed in [15, 19], in distinct scenarios).

Table 3 summarizes the results of our experiments. While the vector-authority ranking provides a gain in average precision of 27%, the vector-hub ranking yields a gain of 25%. Most interesting, the vector-hub-authority ranking leads to a combined gain in average precision of 59%.

Case	Ranking	Vector	Hub	Authority	$P(d_j q)$
1	Vector	yes	no	no	$\eta \times R_{jq}$
2	Hub	no	yes	no	$\eta \times H_{jq}$
3	Authority	no	no	yes	$\eta \times L_{jq}$
4	Vector-Hub	yes	yes	no	$\eta \times [1 - (1 - R_{jq}) \times (1 - H_{jq})]$
5	Vector-Authority	yes	no	yes	$\eta \times [1 - (1 - R_{jq}) \times (1 - L_{jq})]$
6	Vector-Hub-Authority	yes	yes	yes	$\eta \times [1 - (1 - R_{jq}) \times (1 - H_{jq}) \times (1 - L_{jq})]$

Table 1: Alternative rankings modeled in our belief network model.

Number of Pages	Number of Keywords	Average Number of Words per Page	Number of Example Queries	Average Number of Words per Query	Average Number of Pages per Query Pool	Average Number of Relevant Pages per Query Pool
3,027,540	3,465,910	512	20	1.6	38.15	17.05

Table 2: Characteristics of the database.

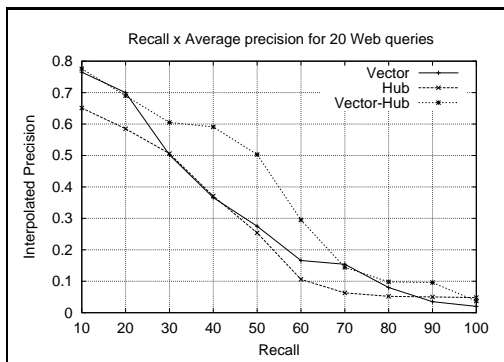


Figure 7: Average precision figures for vector, hub, and vector-hub network rankings.

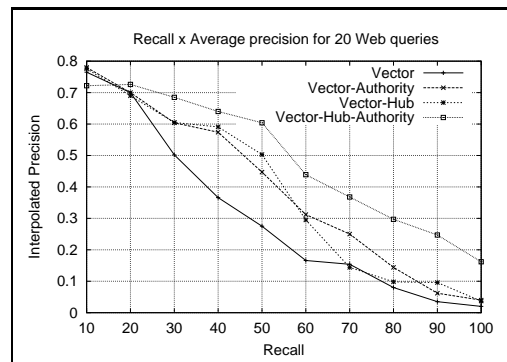


Figure 8: Average precision figures for vector, vector-hub, vector-authority, and vector-hub-authority network rankings.

8 Conclusions

We have described an information retrieval model that combines content-based with link-based pieces of evidence. The model was designed using Bayesian network theory, which provides powerful mechanisms to model the information retrieval problem, specially when distinct sources of evidence are available. The experiments have shown that this combination yields better retrieval performance without requiring any extra information from the users at query time. The combination of hub, authority and content-based evidential information in a single ranking produced an average gain of 59% when compared with the results of the vector space model. Also, it is shown that both hub and authority values are important when ranking Web documents with regard to a user query, and that their combination is better than the use of each of them in isolation.

References

- [1] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, Essex, England, 1999. 513 pages.
- [2] K. Bharat and M. R. Henzinger. Improved algorithms for topic distillation in a hyperlinked environment. In *Proc. of the 21st ACM SIGIR Con-*

ference on Research and Development in Information Retrieval, Distributed Retrieval, pages 104–111, 1998.

- [3] J. Bichteler and E. A. Eaton III. The combined use of bibliographic coupling and cocitation for document retrieval. *Journal of the American Society for Information Science*, 31(7):278–282, 1980.
- [4] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Proc. of the 7th International World Wide Web Conference (WWW7)*, pages 107–117, Brisbane, Australia, 1998.
- [5] S. Chakrabarti, B. Dom, P. Raghavan, S. Rajagopalan, D. Gibson, and J. Kleinberg. Automatic resource compilation by analyzing hyperlink structure and associated text. In *Proc. of the 7th International World Wide Web Conference (WWW7)*, pages 65–74, Brisbane, Australia, 1998.
- [6] A. da Silva, E. Veloso, P. Golgher, B. Ribeiro-Neto, A. Laender, and N. Ziviani. Cobweb - a crawler for the brazilian web. In *Proc of the String Processing and Information Retrieval (SPIRE'99)*, pages 184–191, Cancun, Mexico, 1999.
- [7] W. Frakes and R. Baeza-Yates, editors. *Information*

Average Precision and Gains							
Recall	Vector	Vector-authority	Gain	Vector-hub	Gain	Vector-hub-authority	Gain
10%	0.765	0.780	+1%	0.776	+1%	0.722	-5%
20%	0.700	0.700	+0%	0.690	-1%	0.726	+3%
30%	0.502	0.604	+20%	0.605	+20%	0.685	+36%
40%	0.366	0.574	+56%	0.591	+61%	0.640	+74%
50%	0.275	0.447	+62%	0.503	+82%	0.604	+119%
60%	0.166	0.312	+87%	0.295	+77%	0.439	+164%
70%	0.154	0.250	+62%	0.144	-6%	0.368	+138%
80%	0.080	0.144	+79%	0.098	+22%	0.297	+271%
90%	0.035	0.062	+77%	0.096	+174%	0.247	+605%
100%	0.020	0.040	+100%	0.037	+84%	0.162	+710%
Average	0.306	0.391	+27%	0.384	+25%	0.489	+59%

Table 3: Average precision figures for the vector, vector-authority, vector-hub, and vector-hub-authority network rankings.

- Retrieval: Data Structures & Algorithms*. Prentice Hall, Upper Saddle River, NJ, 1992. 504 pages.
- [8] A. Griffiths, H. C. Luckhurst, and P. Willett. Using interdocument similarity information in document retrieval systems. *Journal of the American Society for Information Science*, 37:3–11, 1986.
- [9] D. Hawking, N. Craswell, and P. Thistlewaste. Overview of trec-7 very large collection track. In *Proc. of the Seventh Text Retrieval Conference - (TREC-7)*, pages 91–104, Gaithersburg, Maryland, 1998. National Institute of Standards and Technology.
- [10] D. Hawking, N. Craswell, P. Thistlewaste, and D. Harman. Results and challenges in web search evaluation. In *Proc. of the 8th International World Wide Web Conference (WWW8)*, Toronto, Canada, 1999. Elsevier Science.
- [11] M. Jansen, A. Spink, J. Bateman, and T. Saracevic. Real life information retrieval: A study of user queries on the web. *ACM SIGIR Forum*, 32(1):5–17, 1998.
- [12] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. In *Proc. of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 668–677, San Francisco, California, 1998.
- [13] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988. 552 pages.
- [14] B. Ribeiro-Neto, I. Silva, and R. Muntz. Bayesian network models for ir. In: *Soft Computing in Information Retrieval: Techniques and Applications*, F. Crestani and G. Pasi editors, Springer Verlag. To appear.
- [15] B. Ribeiro (Ribeiro-Neto) and R. Muntz. A belief network model for ir. In *Proc. of the 19th ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 253–260, Zurich, Switzerland, 1996.
- [16] G. Salton and M. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, NY, 1983. 448 pages.
- [17] Gerard Salton. *Automatic Information Organization and Retrieval*. McGraw-Hill, New York, NY, 1968.
- [18] H. G. Small and M. E. D. Koenig. Journal clustering using a bibliographic coupling method. *Information Processing & Management*, 13:277–288, 1977.
- [19] H. Turtle and W. Croft. Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems*, 9(3):187–222, 1991.
- [20] I. Witten, A. Moffat, and T. Bell. *Managing Gigabytes: Compressing and Indexing Documents and Images*. Morgan Kaufmann, San Francisco, CA, 2nd edition, 1999. 519 pages.
- [21] S. Wong and Y. Yao. A probabilistic inference model for information retrieval. *Information Systems*, 16(3):301–321, 1991.